De quoi parle ce Tweet? Résumer Wikipédia pour contextualiser des microblogs

Romain Deveaud, Florian Boudin

Titre ouvrage :

De quoi parle ce Tweet ? Résumer Wikipédia pour contextualiser des microblogs

Nom de(s) auteur(s) :

— Romain Deveaud, University of Glasgow, UK, (romain.deveaud@glasgow.ac.uk)

— Florian Boudin, LINA - UMR CNRS 6241, Université de Nantes, France, (florian.boudin@univ-nantes.fr)

Résumé :

Les réseaux sociaux sont au centre des communications sur internet et une grande partie des échanges communautaires se fait à travers eux. Parmi eux, l'apparition de Twitter a donné lieu à la création d'un nouveau type de partage d'informations où les messages sont limités à 140 caractères. Les utilisateurs de ce réseau s'expriment donc succinctement, souvent en temps réel à partir d'un smartphone, et la teneur des messages peut parfois être difficile à comprendre sans contexte. Nous proposons dans cet article une méthode permettant de contextualiser automatiquement des Tweets en utilisant des informations provenant directement de l'encyclopédie en ligne Wikipédia, avec comme but final de répondre à la question : « De quoi parle ce Tweet ? ». Nous traitons ce problème comme une approche de résumé automatique où le texte à résumer est composé d'articles Wikipédia liés aux différentes informations exprimées dans un Tweet. Nous explorons l'influence de différentes méthodes de recherche d'articles liés aux Tweets, ainsi que de plusieurs caractéristiques utiles pour la sélection des phrases formant le contexte. Nous évaluons notre approche en utilisant la collection de la tâche Tweet Contextualization d'INEX 2012 et donnons un aperçu sur ce qui caractérise une phrase importante pour déterminer le contexte d'un Tweet.

Mots clés :

Contexte informatif, résumé par extraction, Twitter, Wikipédia

Page des mentions légales

1 Introduction

La grande démocratisation de l'accès à internet et l'avènement des smartphones ont changé le paysage virtuel et la nature des échanges entre les personnes. L'information n'attend plus forcément d'être trouvée par quelqu'un ayant un besoin précis, elle vient directement à nous. Au centre de ce phénomène, les réseaux sociaux sont un média privilégié pour la diffusion de contenu à grande échelle [2]. Les utilisateurs sont reliés par des connections de natures diverses (professionnelles, personnelles, publicitaires...) et s'échangent des informations en temps réel sur le monde qui les entoure. Twitter fait partie de ces réseaux sociaux et favorise des échanges de messages très courts. Quand il se connecte à Twitter, l'utilisateur doit répondre à la question « Quoi de neuf ? ». La réponse à cette question doit faire moins de 140 caractères et est appelée un *Tweet*. De par sa taille, un Tweet est naturellement ambigu et souvent sous-spécifié, ce qui peut rendre la compréhen-sion compliquée pour une personne ne possédant pas le contexte approprié. Ce contexte peut être formé de phrases récupérées sur le Web (ou toute autre source) et réunies afin d'éclairer les lecteurs d'un Tweet sur sa nature et sur les concepts informatifs mis en jeu.

Des travaux similaires ont cherché à comprendre les informations liées aux Tweets et à découvrir plusieurs formes de contextes, aussi en extrayant leurs entités nommées [21], en liant les Tweets à une base de connaissance [16] ou les regroupant sous forme de groupements thématiques [14]. Toutes ces approches présentent néanmoins le désavantage de laisser l'utilisateur chercher par lui-même les informations liées aux Tweets qui auront été identifiées. De son côté, l'approche détaillée dans cet article présente directement à l'utilisateur un contexte lisible et informatif, formé à partir de phrases extraites de Wikipédia. Notre approche de la contextualisation met en jeu successivement des techniques de Recherche d'Information (RI) et de résumé automatique. Tout d'abord, nous cherchons à améliorer la compréhension du Tweet en récupérant des articles Wikipédia pertinents par rapport à celui-ci. Ces derniers sont susceptibles de contenir des passages informatifs pour la construction du contexte du Tweet. Ensuite, nous considérons la formation du contexte comme une tâche de résumé automatique multi-document, où il s'agit de résumer les articles Wikipédia retournés.

Nous évaluons les performances de l'approche proposée en utilisant l'ensemble de données issu de la tâche *Tweet Contextualization* d'INEX [22], qui propose un cadre expérimental permettant d'évaluer la contextualisation de Tweets réalisée à l'aide de phrases issues de Wikipédia. Les résultats expérimentaux montrent que notre approche est robuste, et obtient les seconds meilleurs résultats sur les données de l'année 2012 et les meilleurs résultats sur les données de l'année 2013. De plus, les organisateurs de cette tâche ont également évalué manuellement la lisibilité des contextes proposés par les participants. Nous reportons ces résultats de lisibilité concernant l'édition 2013, où notre approche obtient les seconds meilleurs résultats.

2 Travaux précédents

Le problème de contextualisation de messages courts est émergent et se situe aux confluents de la Recherche d'Information (RI) ciblée et du résumé automatique. Twitter est habituellement utilisé comme une source d'informations récentes [23], à partir de laquelle des articles journalistiques traitant de sujets « chauds » peuvent être recommandés [6]. Différentes approches ont proposé des solutions pour pallier au problème d'ambiguité inhérent aux Tweets,

soit en détectant les entités nommées présentes [21], en liant les Tweets à une base de connaissances [16] ou en formant des groupement thématiques [14].

Ces approches souffrent néanmoins d'un manque d'accessibilité et de lisibilité pour les utilisateurs. La tâche *Tweet Contextualization* de la campagne d'évaluation INEX 2012 est la première à proposer un cadre d'évaluation formel pour ce type de problématique, et a été suivie par de nombreux participants. Ceux-ci doivent, pour un ensemble de Tweets donnés, fournir des contextes constitués de phrases issues d'une édition fixée de Wikipédia. Les approches notables se sont basées sur de la recherche et de l'extraction de passages en utilisant des algorithmes de modélisation thématique [13], ou sur des traitements syntaxiques et sur une reconnaissance des entités nommées dans les phrases candidates [11]. Notre approche est par contre la seule à combiner des techniques de RI, pour découvrir le contexte du Tweet, avec des techniques de résumé automatique, pour extraire les informations les plus saillantes sur ce contexte.

Dans la même veine, une nouvelle tâche de *Temporal Summarization* a fait son apparition à TREC lors de l'édition 2013 [1]. Le but est ici de produire des résumés évoquant des événements majeurs (ouragans, élections...) qui de plus devront être cohérents d'un point de vue chronologique.

Au cours de la dernière décennie, de nombreux chercheurs se sont penchés sur la problématique du résumé automatique. La quasi-totalité des approches proposées recourent à des méthodes d'extraction où il s'agit d'identifier les unités textuelles, le plus souvent des phrases, les plus importantes des documents. Les phrases les plus saillantes sont ensuite assemblées pour générer le résumé.

De nombreuses méthodes ont été utilisées pour évaluer l'importance des phrases, e.g. [3, 20]. Parmi elles, les méthodes basées sur les modèles de graphes [18] donnent de bons résultats. L'idée est de représenter le texte sous la forme d'un graphe d'unités textuelles (phrases) inter-connectées par des relations de similarité. Des algorithmes d'ordonnancement tels que PAGERANK [19] sont ensuite utilisés pour sélectionner les phrases les plus centrales dans le graphe.

Le résumé automatique orienté [9] est probablement la tâche qui se rapproche le plus de la contextualisation automatique. Il s'agit de générer un résumé répondant à un besoin utilisateur exprimé sous la forme d'une requête. Une grande partie des approches proposées reposent sur des méthodes de résumé automatique existantes et y ajoutent divers critères de pertinence par rapport à la requête, e.g. [5]. Parmi les différentes méthodes utilisées pour estimer la pertinence des phrases, plusieurs modèles issus de la RI donnent de bons résultats [25].

3 Recherche de phrases candidates contextuelles issues de Wikipédia

Pour tout le reste de cet article, nous définissons le contexte d'un Tweet comme étant un ensemble de phrases issues de Wikipédia, dont la longueur total n'excède pas 500 mots afin de pouvoir être affiché sur un écran de téléphone portable. Notre approche est constituée de deux étapes principales : une recherche d'articles Wikipédia contenant des informations liées au Tweet (extension du Tweet), puis un résumé de ces articles permettant d'extraire les phrases les plus saillante afin de construire le contexte. Nous présentons dans cette section les différentes étapes que nous mettons en jeu pour former ce contexte, qui sont illustrées par

la Figure 1.

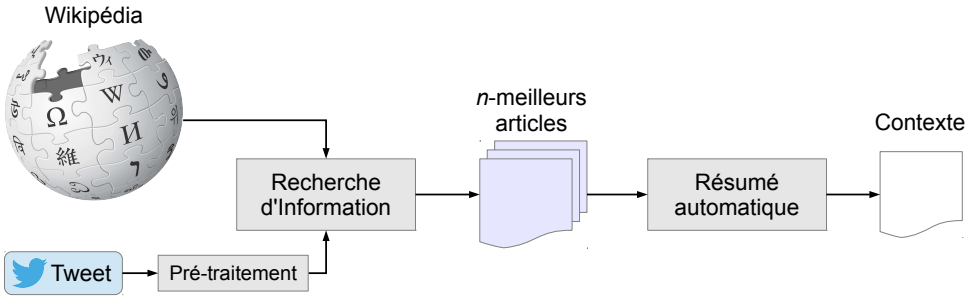

FIGURE 1 – *Méthodologie de contextualisation d'un Tweet à partir de Wikipédia.*

3.1 Interprétation des #HashTags et formatage des Tweets

La première étape que nous effectuons consiste à appliquer un ensemble de pré-traitements aux Tweets. Il s'agit de formater le contenu de ces derniers en vue de l'étape de recherche d'information.

Le symbole #, appelé *hashtag* ou mot-dièse en français, est un caractère utilisé pour préfixer les mots-clés dans les Tweets. Les *hashtags* permettent de catégoriser les Tweets en sujets de discussion. Les utilisateurs emploient les *hashtags* avant un mot-clé ou une phrase pertinente (sans espace) de leurs Tweets. Ils agissent comme un moyen de catégorisation et d'étiquetage et sont ainsi des marqueurs d'informations importantes directement fournis par l'auteur. Ce sont également des marqueurs importants diverses tâches, comme la détection des sentiments [10]. Il semble donc logique de privilégier leur utilisation dans le cadre d'une récupération d'articles Wikipédia liés à un Tweet.

La principale difficulté avec l'utilisation des *hashtags* vient du fait qu'ils sont pour la plupart composés de plusieurs mots concaténés. Pour résoudre ce problème, nous avons utilisé un algorithme de segmentation automatique de mots basé sur celui présenté dans le chapitre « Natural Language Corpus Data » du livre « Beautiful Data » [24]. Nous calculons le découpage le plus probable d'un *hashtag* à l'aide des probabilités d'apparition d'unigrammes et de bigrammes au sein du corpus Bing N-Gram [1]. Ainsi, chaque *hashtag* présent dans le Tweet initial est remplacé par sa version découpée.

Twitter étant un réseau social, l'interaction entre les utilisateurs est au centre de son fonctionnement. Ainsi, un Tweet peut contenir différentes mentions destinées à d'autres personnes, comme par exemple une réponse ou un retweet. Un Tweet réponse commence par un @ suivi du pseudonyme d'un (ou plusieurs) utilisateur(s). Cela permet notamment de créer une discussion spontanée entre plusieurs personnes. Les expériences préliminaires que nous avons menées ont montré que les pseudonymes n'avaient aucune influence positive sur la récupération d'articles Wikipédia, et pouvaient dégrader les résultats dans certains cas. Nous avons donc fait le choix de les supprimer. Quant au retweet, il consiste à reposter le Tweet d'une autre personne. Parfois les utilisateurs tapent *RT* au début d'un Tweet pour indiquer qu'ils repostent le contenu d'un autre utilisateur. Ce n'est pas une commande ou une fonction officielle de Twitter, mais cela signifie qu'ils citent le Tweet d'un autre utilisateur. Néanmoins ces différentes mentions n'apportent rien au contenu informatif du Tweet, nous supprimons

1. http://web-ngram.research.microsoft.com/info/

donc simplement les deux lettres *RT* lorsqu'elle apparaissent. Les mots outils sont également supprimés en utilisant la liste standard INQUERY fournie avec le système de recherche d'information Indri [2]. La sortie finale de cette étape de formatage est un Tweet nettoyé, sans mots-outils, ni *hashtags* collés, ni mentions inutiles.

3.2 Recherche d'articles Wikipédia

La sélection d'articles Wikipédia apportant des informations contextuelles par rapport à un Tweet est une étape cruciale pour trouver les phrases qui vont former le contexte. Nous présentons dans cette section les différentes méthodes de recherche documentaire que nous utilisons dans nos expériences.

3.2.1 Modèle de base

L'une des approches standard de la recherche d'information par modèle de langue se fait avec un modèle de vraisemblance de la requête. Ce modèle mesure la probabilité que la requête puisse être générée à partir d'un document donné, ainsi les documents sont ordonnés en se basant sur cette probabilité. Soit θ_D le modèle de langue estimé en se basant sur un document D, le score d'appariement entre D et une requête $\mathcal{T}$ est défini par la probabilité conditionnelle suivante :

$$P(\mathcal{T}|\theta_D) = \prod_{t \in \mathcal{T}} f_T(t, D) \tag{1}$$

Un des points importants dans le paramétrage des approches par modèle de langue est le lissage des probabilités nulles. Dans ce travail, θ_D est lissé en utilisant le lissage de Dirichlet [26], on a donc :

$$f_T(t, D) = \frac{c(t, D) + \mu \cdot P(t|\mathcal{C})}{|D| + \mu} \tag{2}$$

où $c(t, D)$ est le nombre d'occurrences du mot t dans le document D. $\mathcal{C}$ représente la collection de documents et μ est le paramètre du lissage de Dirichlet (nous fixons $\mu = 2500$ tout au long de cet article).

Une des limitations évidente de l'approche par unigramme est qu'elle ne tient pas compte des dépendances ou des relations qu'il peut y avoir entre deux termes adjacents dans la requête. Le modèle MRF (Markov Random Field) [17] est une généralisation de l'approche par modèle de langue et résoud spécifiquement ce problème. L'intuition derrière ce modèle est que des mots adjacents de la requête sont susceptibles de se retrouver proches dans les documents pertinents. Trois différents types de dépendances sont considérés :

1. l'indépendance des termes de la requête (ce qui revient à un modèle de langue standard prenant en compte uniquement les unigrammes),
2. l'apparition exacte de bigrammes de la requête,
3. et l'apparition de bigrammes de la requête dans un ordre non défini au sein d'une fenêtre de mots.

Le modèle propose deux fonctions supplémentaires, analogue à f_T, pour deux autres types de dépendances qui agissent sur les bigrammes de la requête. La fonction $f_O(q_i, q_{i+1}, D)$ considère la correspondance exacte de deux mots adjacents de la requête. Elle est dénotée par

2. `http://www.lemurproject.org/indri`

l'indice O. La seconde est dénotée par l'indice U et considère la correspondance non ordonnée de deux mots au sein d'une fenêtre de 8 unités lexicales.

Finalement, le score d'un article Wikipédia D par rapport à un Tweet formaté $\mathcal{T}$ est donné par la fonction suivante :

$$s_{MRF}(\mathcal{T}, D) = \lambda_T \prod_{t \in \mathcal{T}} f_T(t, D) + \lambda_O \prod_{i=1}^{|\mathcal{T}|-1} f_O(t_i, t_{i+1}, D) + \lambda_U \prod_{i=1}^{|\mathcal{T}|-1} f_U(t_i, t_{i+1}, D) \quad (3)$$

où λ_T, λ_O et λ_U sont des paramètres libres dont la somme est égale à 1. Dans nos expériences nous fixons ces paramètres en suivant les recommandations des auteurs ($\lambda_T = 0,85$, $\lambda_O = 0,10$ et $\lambda_U = 0,05$) [17].

3.2.2 Intégration de *hashtags*

Nous avons vu précédemment que les *hashtags* pouvaient être considérés comme des requêtes courtes, une sorte d'abréviation du Tweet. Considérons le Tweet $\mathcal{T}$ suivant :

> « #physician #jobs Colon Rectal Surgeon Needed Locums or Perm http://t.co/nnkchYhSKz »

Le sujet principal est correctement représenté par un ensemble de *hashtags* $H_\mathcal{T}$ = {"physician", "jobs"}. Le reste des mots du Tweet sert à spécifier les détails de l'emploi en question. Nous pouvons ainsi considérer cet ensemble de *hashtags* comme une simplification du Tweet ou encore une expression des informations les plus importantes. Un parallèle peut également être fait avec les *topics* de TREC qui sont traditionnellement composés d'une requête courte (2 à 5 mots-clés) et d'une description plus détaillée du besoin d'information (pouvant comprendre plusieurs phrases).

Nous introduisons donc les *hashtags* de façon explicite dans la fonction de score des articles Wikipédia de notre système. Soient un Tweet $\mathcal{T}$ et ses *hashtags* $H_\mathcal{T}$, le score d'un article Wikipédia D est donné par :

$$s(\mathcal{T}, H_\mathcal{T}, D) = \alpha\, s_{MRF}(H_\mathcal{T}, D) + (1 - \alpha)\, s_{MRF}(\mathcal{T}, D) \quad (4)$$

Le paramètre α permet de calibrer l'importance donnée aux *hashtags* seuls dans la requête. Nous nous plaçons dans le cadre d'une contextualisation en temps réel, et la nature très hétérogène des Tweets ne nous semble pas adaptée pour effectuer un apprentissage *a priori* de ce paramètre. De plus, les *hashtags* peuvent avoir une utilité parfois très limitée voire nulle, comme dans l'exemple suivant :

> « U Just Heard "Hard To Believe" by @andydavis on the @mtv Teen Mom 2 Finale go 2 http://t.co/iwb2JuL8 for info #iheareditonMTV »

Dans ce cas-ci, « I heard it on MTV » est une phrase d'accroche de type publicitaire et n'apporte rien pour la compréhension du Tweet. L'importance des *hashtags* est donc elle aussi contextuelle et dépend de leur pouvoir discriminant. Nous choisissons d'estimer ce pouvoir discriminant en calculant un score de clarté [7]. Ce score est en réalité la divergence

de Kullback-Leibler entre le modèle de langue de l'ensemble de *hashtags* et le modèle de langue de la collection $\mathcal{C}$ d'articles Wikipédia :

$$\alpha = \sum_{w \in V} P(w|H_{\mathcal{T}}) \log \frac{P(w|H_{\mathcal{T}})}{P(w|\mathcal{C})} \tag{5}$$

où V représente le vocabulaire. Le modèle de langue des *hashtags* est estimé par retour de pertinence simulé :

$$P(w|H_{\mathcal{T}}) = \sum_{D \in R} P(w|D)P(D|H_{\mathcal{T}}) \tag{6}$$

Nous utilisons pour cela une approche standard de retour de pertinence simulé. Celle-ci consiste à récupérer l'ensemble R constitué des 5 premiers documents [3] de la collection $\mathcal{C}$ renvoyés pour la requête $H_{\mathcal{T}}$. Dans le modèle des *hashtags*, la probabilité $P(D|H_{\mathcal{T}})$ est estimée en appliquant le théorème de Bayes : $P(D|H_{\mathcal{T}}) = \frac{P(H_{\mathcal{T}}|D)P(D)}{P(H_{\mathcal{T}})}$, où la probabilité $P(D)$ est égale à zéro pour les documents qui ne contiennent aucun mot de la requête et où la probabilité $P(H_{\mathcal{T}})$ est constante (donc ignorée). Plus les documents utilisés pour estimer le modèle de langue des *hashtags* sont homogènes, plus la divergence de Kullback-Leibler augmente. Ainsi le paramètre α permet de quantifier à quel point les *hashtags* sont précis et à quel point ils permettent de sélectionner des documents distincts du reste de la collection.

Seuls 23% des Tweets utilisés dans l'évaluation officielle de la tâche *Tweet Contextualization* d'INEX 2012 contiennent des *hashtags*. Lorsqu'il n'y en a pas, nous fixons logiquement $\alpha = 0$ dans l'équation (4).

4 Choix des phrases et formation du contexte

Pour un Tweet donné, nous sélectionnons les n articles Wikipédia les plus pertinents selon l'équation 4. Chaque article est découpé en phrases en utilisant la méthode PUNKT de détection de changement de phrases mise en œuvre dans la boîte à outils `nltk` [4].

Dans ce travail nous fixons $n = 5$, et toutes les phrases de ces articles sont considérées comme des phrases candidates. Nous calculons ensuite différentes caractéristiques pour chacune de ces phrases qui nous permettrons de les classer et, ainsi, de former le contexte. Nous détaillons ces caractéristiques dans la section ci-dessous.

Pour pouvoir être compréhensible dans un cas d'utilisation mobile (sur un *smartphone* par exemple), le contexte doit avoir une taille limitée. Les recommandations de la tâche *Tweet Contextualization* d'INEX fixent la taille limite du contexte à 500 mots. Dans cette section, nous présentons la méthode que nous utilisons pour sélectionner les phrases candidates les plus pertinentes et générer le contexte.

4.1 Caractéristiques des phrases

Plusieurs caractéristiques entrent en compte lors de la sélection des phrases candidates. Ces dernières peuvent être regroupées en quatre catégories :

3. Nous avons expérimenté plusieurs nombres de documents sur les données d'entraînement, et nous n'avons observé que peu de variance dans l'estimation de α.

4. `http://nltk.org/`

1. Importance de la phrase vis-à-vis du document d'où elle provient
2. Pertinence de la phrase par rapport au Tweet (y compris les *hashtags*)
3. Pertinence de la phrase par rapport à une page web dont l'URL est dans le Tweet
4. Pertinence du document d'où provient la phrase par rapport au Tweet

Nous avons spécifiquement choisi ces types de caractéristiques car elles sont communément utilisées dans le domaine du résumé automatique. Bien qu'un très grand nombre d'autres caractéristiques pourrait être calculé, nous laissons ces évolutions pour de futurs travaux. Nous détaillons et justifions dans cette section le calcul des différentes caractéristiques que nous utilisons ensuite pour ordonner les phrases par importance et former le contexte. Nous rappelons quelques notations déjà utilisées dans cet article et nous en introduisons de nouvelles dans le tableau suivant :

$\mathcal{T}$	un tweet nettoyé
$H_{\mathcal{T}}$	les *hashtags* du Tweet $\mathcal{T}$
$U_{\mathcal{T}}$	l'URL présente dans le Tweet $\mathcal{T}$
S	une phrase candidate

Les caractéristiques décrites ci-dessous sont largement basées sur le calcul de mesures de recouvrement et de similarité cosinus entre une phrase candidate $S = \{m_1, m_2, ..., m_i\}$ et un Tweet $\mathcal{T} = \{m_1, m_2, ..., m_j\}$. Soit $| \bullet |$ le cardinal de l'ensemble $\bullet$, le recouvrement en mots est donné par :

$$recouvrement(\mathcal{T}, S) = \frac{|S \cap \mathcal{T}|}{min(|S|, |\mathcal{T}|)} \tag{7}$$

Aussi, soient $\vec{S}$ et $\vec{\mathcal{T}}$ les représentations vectorielles de S et $\mathcal{T}$, et $||\vec{\bullet}||$ la norme du vecteur $\vec{\bullet}$, la similarité cosinus est donnée par :

$$cosinus(\mathcal{T}, S) = \frac{\vec{S} \cdot \vec{\mathcal{T}}}{\sqrt{||\vec{S}||\ ||\vec{\mathcal{T}}||}} \tag{8}$$

Les mesures décrites précédemment sont calculées à partir des représenta-tions lexicales nettoyées des phrases et des Tweets. Nous supprimons les mots outils et appliquons le méthode de racinisation (*stemming*) des mots de Porter, et ce uniquement afin de calculer les différentes caractéristiques. La représentation des Tweets utilisée pour effectuer la recherche d'article Wikipédia n'utilise pas de racinisation.

4.1.1 Importance de la phrase dans le document

L'importance d'une phrase par rapport au document dans lequel elle apparaît est estimée avec la méthode TextRank [18]. Chaque document est représenté sous la forme d'un graphe pondéré non dirigé G dans lequel les noeuds V correspondent aux phrases, et les arêtes E sont définies en fonction d'une mesure de similarité. Cette mesure détermine le nombre de mots communs entre les deux phrases, les mots outils ayant été au préalable supprimés et les mots restants *stemmés* avec l'algorithme de Porter. Pour éviter de favoriser les phrases longues, cette valeur est normalisée par les longueurs des phrases. Soit $\#(m, S)$ le nombre

d'occurrences du mot m dans la phrase S, la similarité entre les phrases S_i et S_j est définie par :

$$\textit{Sim}(S_i, S_j) = \frac{\sum_{m \in S_i \cup S_j} \#(m, S_i) + \#(m, S_j)}{\log(|S_i|) + \log(|S_j|)} \tag{9}$$

L'importance d'une phrase est évaluée en tenant compte de l'intégralité du graphe. Nous utilisons une adaptation de l'algorithme PAGERANK [19] qui inclut les poids des arêtes. Le score de chaque sommet V_i est calculé itérati-vement jusqu'à la convergence par :

$$p(V_i) = (1-d) + d \times \sum_{V_j \in \text{voisins}(V_i)} \frac{Sim(S_i, S_j)}{\sum_{V_k \in \text{voisins}(V_j)} Sim(S_k, S_i)} \, p(V_j) \tag{10}$$

où d est un « facteur d'amortissement » (typiquement dans l'intervalle $[0.8, 0.9]$) et $\text{voisins}(V_i)$ représente l'ensemble des nœuds connectés à V_i. Le score de la phrase S correspond au score du nœud qui la représente dans le graphe.

$$c_1 = p(S)$$

4.1.2 Pertinence de la phrase par rapport au Tweet

Intuitivement, les indicateurs de pertinence devraient être les plus importants pour sélectionner des phrases donnant des informations contextuelles par rapport au Tweet. Le recouvrement et la similarité cosinus entre un Tweet $\mathcal{T}$ et une phrase candidate S sont les premières caractéristiques que nous avons mis en place.

$$c_2 = recouvrement(\mathcal{T}, S) \qquad c_3 = cosinus(\mathcal{T}, S)$$

Tout en gardant la logique de l'utilisation des *hashtags*, nous calculons le recouvrement et la similarité cosinus entre chaque phrase et l'ensemble des hashtags du Tweet.

$$c_4 = recouvrement(H_{\mathcal{T}}, S) \qquad c_5 = cosinus(H_{\mathcal{T}}, S)$$

4.1.3 Pertinence de la phrase par rapport à une page web

Les Tweets contiennent parfois des URLs, liens pointant vers des pages web porteuses d'informations contextuelles. Nous utilisons le même type de mesure que précédemment et nous calculons ainsi le recouvrement et la similarité cosinus entre une phrase candidate et le titre $titre(U_{\mathcal{T}})$ de la page web.

$$c_6 = recouvrement\big(titre(U_{\mathcal{T}}), S\big) \qquad c_7 = cosinus\big(titre(U_{\mathcal{T}}), S\big)$$

De la même façon, nous calculons ces deux mesures entre le contenu entier $page(U_{\mathcal{T}})$ de la page web et une phrase candidate.

$$c_8 = recouvrement\big(page(U_{\mathcal{T}}), S\big) \qquad c_9 = cosinus\big(page(U_{\mathcal{T}}), S\big)$$

4.1.4 Pertinence du document par rapport au Tweet

Les articles Wikipédia à partir desquels les phrases candidates sont extraites ont des importances contextuelles différentes par rapport à un Tweet donné. Ainsi, une phrase provenant d'un article bien classé a plus de chance d'être importante qu'une phrase provenant d'un article mal classé. Pour capturer ce comportement, nous définissons la dernière caractéristique comme étant le score d'un document par rapport à un Tweet et ses *hashtags*, normalisé sur l'ensemble R de tous les documents renvoyés :

$$c_{10} = \frac{s(\mathcal{T}, H_{\mathcal{T}}, D)}{\sum_{D' \in R} s(\mathcal{T}, H_{\mathcal{T}}, D')}$$

4.1.5 Score final d'une phrase candidate

Le score d'importance de chaque phrase candidate est obtenu par la combinaison linéaire des scores des critères présentés ci-dessus.

$$score = \sum_{x} \log(c_x + 1) \tag{11}$$

N'ayant pas d'ensemble d'entraînement conséquent, nous n'avons pas pu pondérer les différentes caractéristiques dans la première version de notre système. Nous présentons néanmoins cette évolution en Section 5.4.

4.2 Génération du contexte

Le contexte d'un Tweet est généré par assemblage des phrases candidates les plus importantes. Il est cependant possible que le contexte ainsi obtenu contienne plusieurs phrases redondantes, ce qui dégrade à la fois sa lisibilité et son contenu informatif. Pour résoudre ce problème, nous ajoutons une étape supplémentaire lors de la génération des contextes.

Nous générons tous les contextes possibles à partir des combinaisons des N phrases ayant les meilleurs scores, en veillant à ce que le nombre total de mots soit optimal (i.e. en dessous du seuil de 500 mots et qu'il soit impossible d'ajouter une autre phrase sans dépasser ce seuil). La valeur N est fixée empiriquement au nombre minimum de phrases de meilleurs scores pour atteindre 500 mots, plus quatre phrases. Le contexte retenu au final est celui possédant le score global le plus élevé, ce score étant calculé comme le produit du score de la diversité du résumé, estimé par le nombre de n-grammes différents, et de la somme des scores des phrases.

Afin d'améliorer la lisibilité du contexte généré, l'ordre original du document est conservé si deux phrases sont extraites à partir d'un même document.

5 Évaluation et discussion

Cette section débute par la description de la collection de test que nous utilisons. Nous présentons ensuite dans la Section 5.2 les résultats de notre méthode de contextualisation en ré-utilisant les données de l'édition 2012 de la tâche Tweet Contextualization d'INEX, et nous analysons dans la Section 5.3 l'importance des différentes caractéristiques dans le processus de sélection des phrases candidates. Nous étendons notre évaluation en reportant et

en analysant dans la Section 5.4 les résultats officiels de notre participation à l'édition 2013 de cette même tâche, incluant notamment l'évaluation de lisibilité (non reproductible).

5.1 Cadre expérimental

Nous utilisons dans un premier temps la collection de test de la tâche *Tweet Contextualization* d'INEX 2012 pour nos expérimentations ainsi que les différentes données mises à disposition par les organisateurs [22]. La collection de documents Wikipédia est basée sur une capture de la version anglaise de l'encyclopédie en ligne datant de Novembre 2011 et comprend 3 691 092 articles. Nous avons indexé cette collection avec le moteur de recherche libre Indri en supprimant les mots-outils présents dans la liste INQUERY. Une racinisation légère des mots est également appliquée par l'algorithme de Krovetz. Celle-ci est différente de la racinisation de Porter évoquée dans la Section 4.1, qui elle n'était appliquées qu'aux phrases avant le calcul des différentes caractéristiques.

La collection de test comprend au total 1126 Tweets pour lesquels un système doit produire un contexte. Cependant, nous n'utilisons que le sous-ensemble de 63 Tweets pour lesquels des jugements de pertinence ont été réalisés. Ces jugements ont été générés par un processus de groupement des dix premières phrases des contextes de tous les participants qui ont ensuite été jugées manuellement par les organisateurs.

Lors de l'édition 2013, les organisateurs ont actualisé la collection Wikipédia en prenant une capture datant de Novembre 2012, et ont proposé un nouvel ensemble de Tweets à contextualiser. Un ensemble de 1721 Tweets a été fourni aux participants, dont 120 ont été annotés et donc évalués. Pour l'édition 2013, les organisateurs ont explicitement introduit un plus grand nombre de Tweets contenant des *hashtags*. Les résultats que nous présentons dans les sections suivantes sont issus d'expériences réalisées *a posteriori* concernant la collection 2012 (Sections 5.2 et 5.3), et d'une participation à la compétition concernant la collection 2013 (Section 5.4).

La mesure d'évaluation développée pour cette tâche ne prend pas en compte les exemples négatifs, seules les phrases jugées pertinentes ont été conservées. Les jugements sont donc un ensemble de phrases directement issues de Wikipédia et jugées pertinentes par les organisateurs en fonction de leur importance contextuelle par rapport à un Tweet. Certains Tweets peuvent ainsi avoir un contexte de référence composé d'un grand nombre de phrases, tandis que d'autres peuvent en avoir un nombre très réduit. Ces différences de taille ainsi que le fait qu'une seule référence soit disponible pour chaque Tweet empêchent l'utilisation de la mesure classique ROUGE [15] pour l'évaluation des contextes. Les organisateurs ont donc proposé une mesure d'évaluation qui calcule une divergence entre le contexte produit et les phrases jugées pertinentes [4, 22]. Elle peut prendre en compte des unigrammes stricts, des bigrammes ou des bigrammes avec possibilité d'insertion. La mesure principale utilisée pour départager les systèmes est la troisième (« Bigrammes à trous »).

5.2 Résultats de contextualisation

Nous reportons dans le tableau 1 les résultats de contextualisation pour trois méthodes de recherche d'articles Wikipédia présentées dans la section 3 : l'approche standard par modèle de langue pour la RI (équation 1, notée **QL**), l'approche **MRF** (équation 3) et l'approche mixant MRF pour le Tweet et pour ses *hashtags* (équation 4, notée **MRFH**). Les scores étant calculés en tant que divergences, les scores les plus bas correspondent aux systèmes les plus

performants.

	Unigrammes	Bigrammes	Bigrammes à trous
QL	0.7967	0.8923	0.8940
MRF	0.7883	0.8851	0.8865
MRFH	**0.7872**	**0.8815**	**0.8839**
1^{er} INEX 2012	0.7734	0.8616	0.8623
3^{e} INEX 2012	0.7909	0.8920	0.8938

TABLE 1 – *Résultats de contextualisation pour les 3 différents algorithmes de RI et l'ensemble des caractéristiques pour l'attribution des scores.*

Nous remarquons que les résultats sont relativement proches et qu'il n'y a pas de différence significative entre les trois approches. Néanmoins l'approche qui considère les *hashtags* dans la fonction de score des documents obtient les meilleurs résultats (avec $p = 0.17$ pour un t-test entre **QL** et **MRFH**). Les faibles différences observées entre les méthodes sont sans doute dues à la relative similarité entre les modèles de RI, même si l'on voit que l'utilisation de *hashtags* améliore sensiblement les scores. Il est néanmoins difficile de tirer des conclusions définitives étant donné que seuls 23% des Tweets utilisés pour l'évaluation contiennent au moins un *hashtag*. Nous reportons pour information les résultats officiels du meilleur système [8] [5], ainsi que les résultats du troisième système (nous étions classés seconds) Ce dernier présente des performances similaires à notre approche de base **QL**. Grâce à l'analyse détaillée de l'influence des différentes caractéristiques proposée en section 5.3, nous avons pu établir que la borne supérieure de notre système était de 0.8824 ce qui est encore loin du meilleur score. Cependant, il n'y a pas de différence statistiquement significative entre notre approche **MRFH** et le meilleur système d'INEX 2012.

Après de longues discussions avec les organisateurs, nous faisons l'hypothèse que cette grande différence de score peut être due à deux biais lors de l'évaluation. Le premier se situe lors de la constitution des jugements : pour chaque Tweet, uniquement les dix premières phrases de chaque système sont considérées pour être ensuite jugées manuellement. Or, un des buts de cette tâche étant la lisibilité, les phrases les plus informatives ne se trouvent pas forcément en début de contexte pour pouvoir favoriser la cohérence globale et l'enchaînement des phrases. Le deuxième biais peut se situer au sein de la mesure d'évaluation elle-même, qui ne possède pas de composante visant à pénaliser les phrases non pertinentes. Ainsi, remplir le contexte avec des phrases très diverses permet d'obtenir des meilleurs scores que de faire attention à ne pas ajouter de phrases dégradant la cohérence du contexte.

Pour illustrer ces biais, nous présentons dans la figure 2 un Tweet ainsi que le contexte produit par notre méthode qui a obtenu un score nul lors de notre évaluation. Or, même si ce contexte n'est à l'évidence pas parfait, il apporte tout de même des informations contextuelles sur le Tweet. On peut en effet apprendre que Van Gogh était un peintre et que "The Starry Night" est une de ses compositions, et dont le style transparaît sur d'autres de ses peintures.

Malgré tout, notre approche présente de bonnes performances qui, comme nous allons le voir dans la Section 5.4, permettent d'obtenir des résultats état-de-l'art pour l'édition 2013 de la tâche Tweet Contextualization. Il est également à noter que si les résultats obtenus par

5. À l'heure où nous écrivons ces lignes, les auteurs n'ont toujours pas publié de descriptif de leur approche nous permettant de faire une comparaison complète.

« Very cool ! An interactive animation of van Gogh's "The Starry Night" http ://t.co/ErJCPObh (thanks @juliaxgulia) »

Vincent van Gogh painted at least 18 paintings of"olive trees", mostly in Saint-Rémy in 1889. The olive tree paintings had special significance for Van Gogh. One painting, "Olive Trees in a Mountainous Landscape (with the Alpilles in the Background)", a complement to "The Starry Night", symbolized the divine. In both "The Starry Night" and his olive tree paintings, Van Gogh used the intense blue of the sky to symbolize the "divine and infinite presence" of Jesus. ...

FIGURE 2 – *Les premières phrases d'un contexte produit par notre méthode pour lequel la mesure d'évaluation a attribué un score nul.*

notre méthode lors de la campagne INEX ne sont pas les meilleurs, notre approche est celle qui apporte le meilleur compromis entre informativité et lisibilité. Néanmoins l'évaluation de lisibilité, qui a été faite manuellement, n'est pas reproductible et le travail présenté dans cet article est différent de celui réalisé pour INEX, nous ne pouvons donc pas reporter de résultats pour le jeu de données de l'année 2012. Nous reportons néanmoins des résultats de lisibilité, obtenus à partir du jeu de données de l'année 2013, en Section 5.4.

5.3 Importance des différentes caractéristiques

En l'état, les caractéristiques calculées pour chacune des phrases candidates ont toutes la même importance dans le score final attribué à une phrase. Nous proposons dans cette section une analyse de leur importance sur les Tweets de l'année 2012. Bien évidemment, les chiffres présentés ici ne nous ont pas servi à paramétrer notre système, et les résultats présentés dans la section précédente ne tiennent pas compte de ces poids.

En principe, nous pourrions utiliser n'importe quelle méthode d'apprentissage pour apprendre les poids optimaux. Ici, nous utilisons un modèle de régression logistique. Ainsi, nous calculons toutes les caractéristiques présentées dans la section 4.1 pour chacune des phrases extraites et nous les lions à leur pertinence $r \in \{0, 1\}$. La variable r peut ainsi être vue comme une mesure de la contribution totale de toutes les caractéristiques utilisées dans le modèle et est habituellement définie comme $r = \bar{w}\bar{x}$. Spécifiquement, $\bar{x}$ est un vecteur de valeurs numériques représentant les caractéristiques, et $\bar{w}$ représente l'ensemble des poids relatifs de chaque caractéristique. Un poids positif signifie que la caractéristique correspondante améliore la probabilité d'obtenir r, un poids négatif signifie qu'elle la dégrade.

Nous pouvons observer dans le tableau 2 que les caractéristiques les plus significatives pour estimer la pertinence d'une phrase ne sont pas ou peu liées au Tweet. En effet, le TextRank ne concerne que l'importance de la phrase par rapport aux autres phrases du document, et le score du document est un score global. Le Tweet n'intervient dans ces cas qu'au moment de la recherche des articles. Le recouvrement et la similarité cosinus entre le Tweet et une phrase sont également des marqueurs de pertinence. Étonnamment, les *hashtags* ont une influence parfois négative et généralement aléatoire, tout comme les titres des pages web pointées par les URLs. Mais comme nous l'avons dit dans la section précédente, les *hashtags* sont très peu nombreux dans les Tweets utilisés pour l'évaluation, ce qui peut expliquer ce comportement aléatoire. Enfin, seule la similarité cosinus entre une phrase et le contenu d'une page web semble être faiblement significative.

Caractéristique	Nom	Valeur (w_x)	Significativité
$c1$	**TextRank**	8.996	$p < 2^{-16}$
$c2$	**Recouvrement Tweet**	2.496	$p = 2.38^{-6}$
$c3$	**Cosine Tweet**	5.849	$p = 4^{-15}$
$c4$	Recouvrement *hashtags*	-2.051	$p = 0.1368$
$c5$	Cosine *hashtags*	0.671	$p = 0.3074$
$c6$	Recouvrement titre URL	1.373	$p = 0.2719$
$c7$	Cosine titre URL	0.788	$p = 0.6287$
$c8$	Recouvrement page URL	0.543	$p = 0.4337$
$c9$	**Cosine page URL**	10.374	$p = 0.0195$
$c10$	**Score document**	0.782	$p < 2^{-16}$

TABLE 2 – *Valeurs optimales des poids des caractéristiques calculées pour les phrases candidates.*

Globalement, une phrase apporte des informations contextuelles par rapport à un Tweet si elle contient les mêmes mots que celui-ci, si elle apparaît dans un document pertinent, et si elle fait partie des phrases les plus importantes de ce dernier.

5.4 Évaluation étendue et lisibilité

Les expériences précédentes, menées sur la collection 2012 de la tâche Tweet Contextualization, nous ont permis d'identifier les caractéristiques saillantes d'une phrase apportant des informations contextuelles par rapport à un Tweet. Dans cette section, nous reportons les résultats de notre participation à l'édition 2013 de cette même tâche. Le corpus d'articles Wikipédia n'a pas changé entre les deux années, mais un nouvel ensemble de Tweets à contextualiser a été fourni. Ceux-ci ont été explicitement sélectionnés pour leur diversité, y compris dans l'utilisation des *hashtags*.

Nous avons mené des expérimentations avec trois variantes du système **MRFH** :

— **MRFH**-all-notrain, qui utilise toutes les caractéristiques présentées en Section 4.1, mais sans prendre en compte les valeurs optimales de leur poids (voir Tableau 2). Cela correspond au modèle **MRFH** évalué dans la section précédente.

— **MRFH**-title-only-notrain, qui lui ne considère que les caractéristiques $c2$ et $c3$, uniquement liées au texte du Tweet.

— **MRFH**-all-train, qui utilise toutes les caractéristiques ainsi que leurs poids appris sur en utilisant les Tweets de la collection 2012.

Le score d'une phrase suivant cette dernière approche est calculé selon :

$$score = \sum_x \log(w_x \cdot c_x + 1) \tag{12}$$

où les poids w_x sont pris directement du Tableau 2.

Nous reportons les résultats officiels des 10 systèmes qui ont obtenu les meilleurs résultats d'informativité pour l'année 2013 dans le Tableau 3. La mesure d'évaluation étant toujours une divergence, les scores les plus bas sont les meilleurs. Nous observons que les trois variantes de notre approche obtiennent respectivement les 1ère, 2nde, et 6ème places, ce qui confirme les résultats prometteurs obtenus dans la section précédente. Un plus grand nombre

	Unigrammes	Bigrammes	Bigrammes à trous
MRFH-all-notrain	**0.8861**	**0.8810**	**0.7820**
MRFH-title-only-notrain	0.8943	0.8908	0.7939
275	0.8969	0.8924	0.8061
273	0.8973	0.8921	0.8004
274	0.8974	0.8922	0.8009
MRFH-all-train	0.8998	0.8969	0.7987
254	0.9242	0.9229	0.8331
276	0.9301	0.9270	0.8169
270	0.9397	0.9365	0.8481
267	0.9468	0.9444	0.8838

TABLE 3 – *Résultats officiels de contextualisation pour l'édition 2013 de la tâche Tweet Contextualisation (10 meilleurs systèmes). Les noms des autres participants ont été anonymisés.*

de *hashtags* ont été introduits dans les Tweets de l'édition 2013, ces résultats confirment donc que leur utilisation comme des éléments à part entière pour récupérer les articles Wikipédia est indispensable.

Nous pouvons faire deux remarques particulièrement intéressantes par rapport aux résultats de nos systèmes. Premièrement, apprendre les poids des caractéristiques s'est trouvé être handicapant pour la sélection des phrases. En effet, les deux systèmes obtenant les meilleurs résultats utilisent des poids identiques pour toutes les caractéristiques. Nous pensons que cela est principalement dû au changement de nature des Tweets entre les deux années. Deuxièmement, nous pouvons voir que le deuxième meilleur système n'utilise que deux caractéristiques pour classer les Tweets, et que celles-ci sont uniquement liées à leur texte. Cela confirme l'importance des caractéristiques $c2$ et $c3$, déjà identifié dans le Tableau 2. De plus, ce résultat est prometteur car il montre que notre modèle de recherche d'article Wikipédia est suffisamment robuste pour effectuer un filtrage de qualité, réduisant alors la nécessité d'un lourd travail de traitement des phrases candidates.

La tâche Tweet Contextualisation évalue également la lisibilité des contextes produits. Néanmoins, c'est une évaluation manuelle qui n'est réalisée qu'une seule fois par des assesseurs, nous n'avions donc pas pu reporter ses résultats pour l'année 2012. Nous reportons cependant l'évaluation de lisibilité de l'année 2013 pour les 10 meilleurs systèmes dans le Tableau 4. La lisibilité est évaluée selon 4 critères [22], qui sont ensuite moyennés :

— Pertinence : juge si les phrases ont un sens au sein de leur contexte (i.e. après avoir lu les autres phrases dans le même contexte).

— Nouveauté : évalue la capacité du contexte à ne pas contenir trop d'informations redondantes.

— Justesse : évalue la résolution d'anaphores au sein du contexte.

— Syntaxe : évalue les problèmes syntaxiques.

Notre système **MRFH**-all-notrain n'obtient pas les meilleurs résultats, mais reste très proche du premier. Il est toutefois intéressant de noter que la méthode heuristique détaillée dans la Section 4.2 permet d'éliminer les phrases redondantes et d'obtenir les meilleurs scores en terme de Nouveauté. L'évaluation de lisibilité semble néanmoins être très dépendante de

	Moyenne	Pertinence	Nouveauté	Justesse	Syntaxe
275	**72.44%**	**76.64%**	67.30%	**74.52%**	**75.50%**
MRFH-all-notrain	72.13%	74.24%	**71.98%**	70.78%	73.62%
274	71.71%	74.66%	68.84%	71.78%	74.50%
273	71.35%	75.52%	67.88%	71.20%	74.96%
MRFH-all-train	69.54%	72.18%	65.48%	70.96%	72.18%
254	67.46%	73.30%	61.52%	68.94%	71.92%
MRFH-title-only-notrain	65.97%	68.36%	64.52%	66.04%	67.34%
276	49.72%	52.08%	45.84%	51.24%	52.08%
267	46.72%	50.54%	40.90%	49.56%	49.70%
270	44.17%	46.84%	41.20%	45.30%	46.00%

TABLE 4 – *Résultats officiels de lisibilité pour l'édition 2013 de la tâche Tweet Contualization (10 meilleurs systèmes).*

l'annotateur. En effet, les trois variantes que nous évaluons se basent sur le même modèle de recherche d'articles Wikipédia et ne diffèrent que sur les phrases sélectionnées. Même lors de cette sélection, la différence entre les valeurs des caracté-ristiques reste assez faibles, ce qui implique que les contextes générés sont relativement similaires (c'est d'ailleurs ce que l'on voit dans le Tableau 3, où nos deux meilleures approches sont très similaires et obtiennent des résultats très proches). Ainsi, il semble assez peu probable que d'aussi grandes différences en lisibilité puissent être observées, à moins que les trois systèmes aient été évalués par des assesseurs différents. Pour les futures éditions de la tâche Tweet Contextualization, nous pensons que les organisateurs pourraient faire évaluer la lisibilité de chaque contexte par plusieurs assesseurs. Le report de chiffres tels que les accords inter-annotateurs pourrait ainsi grandement aider l'analyse de ces résultats.

6 Conclusion

Nous avons présenté dans cet article une première approche pour la contextualisation de messages courts, qui a obtenu respectivement les deuxièmes et premières places lors des éditions 2012 et 2013 de la tâche Tweet Contextualization d'INEX. De plus cette approche a obtenu des résultats de lisibilité très satisfaisants même si elle ne prenait pas explicitement ce critère en compte. Les résultats de nos expériences suggèrent que l'utilisation des *hashtags* présents dans les Tweets aide à la recherche d'articles Wikipédia qui contiennent des phrases apportant des informations contextuelles. En examinant l'influence des différentes caractéristiques calculées sur les phrases, nous avons trouvé que les phrases centrales des articles Wikipédia récupérés en utilisant le Tweet comme requête sont les meilleures candidates pour la génération des contextes. Les mesures de similarité entre les phrases et les Tweets sont également des indicateurs fiables.

Références

[1] ASLAM, J., DIAZ, F., EKSTRAND-ABUEG, M., PAVLU, V., AND SAKAI, T. TREC

2013 Temporal Summrization. In *Proceedings of TREC* (2013).

[2] BAKSHY, E., ROSENN, I., MARLOW, C., AND ADAMIC, L. The role of social networks in information diffusion. In *Proceedings of WWW* (2012), pp. 519–528.

[3] BARZILAY, R., ELHADAD, M., ET AL. Using lexical chains for text summarization. In *Proceedings of the ACL workshop on intelligent scalable text summarization* (1997), vol. 17, pp. 10–17.

[4] BELLOT, P., MORICEAU, V., MOTHE, J., SANJUAN, E., AND TANNIER, X. Evaluation de la contextualisation de tweets. In *Proceedings of CORIA* (2013), pp. 141–148.

[5] BOUDIN, F., EL-BÈZE, M., AND TORRES-MORENO, J.-M. A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization. In *Proceedings of Coling* (2008), pp. 23–26.

[6] CHEN, J., NAIRN, R., NELSON, L., BERNSTEIN, M., AND CHI, E. Short and tweet : experiments on recommending content from information streams. In *Proceedings of CHI* (2010), pp. 1185–1194.

[7] CRONEN-TOWNSEND, S., AND CROFT, W. B. Quantifying query ambiguity. In *Proceedings of HLT* (2002), pp. 104–109.

[8] CROUCH, C. J., CROUCH, D. B., CHITTILLA, S., NAGALLA, S., KULKARNI, S., AND NAWALE, S. The 2012 INEX Snippet and Tweet Contextualization Tasks. In Forner et al. [12].

[9] DANG, H. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference* (2005).

[10] DAVIDOV, D., TSUR, O., AND RAPPOPORT, A. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In *Proceedings of Coling* (2010), pp. 241–249.

[11] ERMAKOVA, L., AND MOTHE, J. IRIT at INEX 2012 : Tweet Contextualization. In Forner et al. [12].

[12] FORNER, P., KARLGREN, J., AND WOMSER-HACKER, C., Eds. *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012* (2012).

[13] GANGULY, D., LEVELING, J., AND JONES, G. J. F. DCU@INEX-2012 : Exploring Sentence Retrieval for Tweet Contextualization. In Forner et al. [12].

[14] GENC, Y., SAKAMOTO, Y., AND NICKERSON, J. V. Discovering Context : Classifying Tweets Through a Semantic Transform Based on Wikipedia. In *Proceedings of FAC* (2011), pp. 484–492.

[15] LIN, C.-Y. ROUGE : A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out : Proceedings of the ACL-04 Workshop* (2004), S. S. Marie-Francine Moens, Ed., ACL, pp. 74–81.

[16] MEIJ, E., WEERKAMP, W., AND DE RIJKE, M. Adding Semantics to Microblog Posts. In *Proceedings of WSDM* (2012), pp. 563–572.

[17] METZLER, D., AND CROFT, W. B. A Markov Random Field Model for Term Dependencies. In *Proceedings of SIGIR* (2005), pp. 472–479.

[18] MIHALCEA, R. Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. In *Proceedings of ACL* (2004), pp. 170–173.

[19] PAGE, L., BRIN, S., MOTWANI, R., AND WINOGRAD, T. The pagerank citation ranking : bringing order to the web.

[20] RADEV, D., JING, H., STYŚ, M., AND TAM, D. Centroid-based summarization of multiple documents. *Information Processing & Management 40* (2004), 919–938.

[21] RITTER, A., CLARK, S., MAUSAM, AND ETZIONI, O. Named Entity Recognition in Tweets : An Experimental Study. In *Proceedings of EMNLP* (2011), pp. 1524–1534.

[22] SANJUAN, E., MORICEAU, V., TANNIER, X., BELLOT, P., AND MOTHE, J. Overview of the INEX 2012 Tweet Contextualization Track. In Forner et al. [12].

[23] SANKARANARAYANAN, J., SAMET, H., TEITLER, B. E., LIEBERMAN, M. D., AND SPERLING, J. TwitterStand : news in tweets. In *Proceedings of GIS* (2009), pp. 42–51.

[24] SEGARAN, T., AND HAMMERBACHER, J. *Beautiful Data : The Stories Behind Elegant Data Solutions*. O'Reilly Media, 2009.

[25] WEI, F., LI, W., LU, Q., AND HE, Y. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *Proceedings of SIGIR* (2008), pp. 283–290.

[26] ZHAI, C., AND LAFFERTY, J. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems 22*, 2 (2004).